International Journal of Advances in Intelligent Informatics
Vol. 4, No. 1, March 2018, pp. 63-69
ISSN 2442-6571

The performance of text similarity algorithms

Didik Dwi Prasetya [a,b,1,*], Aji Prasetya Wibawa [a,2], Tsukasa Hirashima [b,3]

[a] Department of Electrical Engineering, State University of Malang, Indonesia
[b] Graduate School of Engineering, Hiroshima University, Japan
[1] didikdwi@um.ac.id; [2] aji.prasetya.ft@um.ac.id; [3] tsukasa@lel.hiroshima-u.ac.jp
* corresponding author

ARTICLE INFO

Article history
Received February 16, 2018
Revised March 31, 2018
Accepted March 31, 2018

Keywords
Similarity measure
String-based
Corpus-based
Knowledge-based
Text Mining

ABSTRACT

Text similarity measurement compares text with available references to indicate the degree of similarity between those objects. There have been many studies of text similarity and resulting in various approaches and algorithms. This paper investigates four majors text similarity measurements, which include String-based, Corpus-based, Knowledge-based, and Hybrid similarities. The results of the investigation showed that the semantic similarity approach is more rational in finding substantial relationship between texts.

1. Introduction

The recent information problematic issue is the rapid data growth [1]. Text similarity measurement is a text mining approach that could be overcome this overwhelming problem. Finding the similarity between words is a primary stage for sentence, paragraph and document similarities [2]. Text similarity approach may alleviate people on finding relevant information. This is the backbone of successful text mining operations such as searching and information retrieval (IR), text classification, information extraction (IE), document clustering [3], sentiment analysis, machine translation, text summarization, and natural language processing (NLP).

Lexical and semantic similarity words is an essential element of sentence, paragraph and document similarity measurement [2]. Lexical similarity a degree of two given string are similar in its character sequence. While the score is one (1), means the words are 100% lexically identic. In contrast, zero (0) indicates that there is no common word between given strings. On the other hand, semantic similarity represents the likeness among text and document on the basis of their contextual meaning. For example, the pair of "book" and "cook" have a high lexical similarity, but they are not semantically related. The pair of "car" and "wheel" that seems have no lexical similarity, but they are very semantically related as they are automotive-related terms.

Gomaa [2] explained the three main categories of text similarity approach, but did not discuss about the evaluation of algorithms performance. This paper will survey the measurement approaches of lexically and semantically text similarities from the widely used to the recent issues. This study also evaluate the ten most common algorithms that represents each category of text similarity measure.

2. Method

To complete the study of this text similarity, we conducted a performance investigation of text similarity algorithms. In this evaluation, three pairs of texts are used, took from Barron's research [4]. The pairs are, Pair 1 ("book", "cook"); Pair 2 ("car", "wheel"); and Pair 3 ("antique", "ancient").

Refering to the test data, we can see that the texts in first pair represent lexical similarity, while the second pair describe the semantic similarity. The last pair represent that both texts have lexical and semantic similarity. This evaluation involves ten algorithms from four categories of text similarity measure we have describe. To test these algorithms we used several libraries, such as SimMetrics, SoftTFIDF, WS4J, and SEMILAR. In this test, we only focus on the retuned similarity score by each executed algorithm.

3. Results and Discussion

3.1. Text similarity algorithms

Different approaches have been promoted to measure the similarity between one text with another. The method is divided into four major groups, String-based, Corpus-based, Knowledge-based, and Hybrid text similarities; as shown in Fig. 1. These approaches will be detailed in the following subsections.

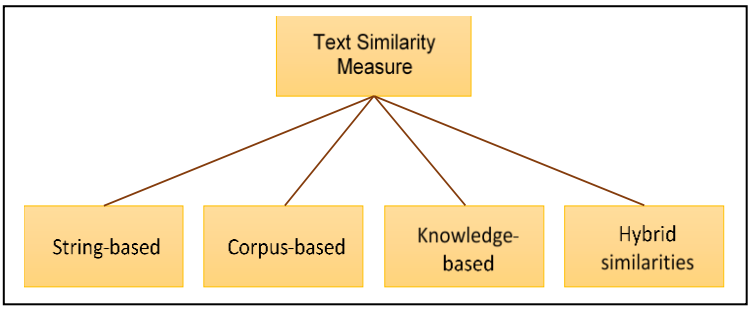

Fig. 1. Four major groups of text similarity methods and algorithms

3.1.1. Categories of text similarity String-based Similarity

String-based similarity is the oldest, simplest yet most popular measurement approach. This measure operates on string sequences and character composition. Two main types of string similarity functions are character-based similarity functions, and token-based similarity functions.

Character-based Similarity is also called sequence-based or edit distance (ED) measurement. It takes two strings of characters and then calculates the edit distance (including insertion, deletion and substitution) between them. Character-based quantifies character similarity between two strings to quantify the similarity, for instance edit distance which is the minimum number of single-character edit operations needed to transform one to another [5]. In another word, two strings are similar if the edit distance minimum operation number is smaller than the given threshold. Some examples of this approach are Hamming distance [6], Levenshtein distance [7]. Damerau-Levenshtein [7], [8], Needleman-Wunsch [9], Longest Common Subsequence [10]. Smith-Waterman [11], Jaro [12], Jaro-Winkler [13], and N-gram [14], [15]. Character-based measure is useful for recognizing typographical errors, but it is useless in recognition of the rearranged terms (e.g. data analyzing and analyzing data) [16]. Edit distance is widely used for string matching approximation to handle the existing data inconsistence [17].

The term-based similarity also known as token-based because it models each string as a set of tokens. The similarity between strings can be assessed by manipulating sets of tokens, such as words. The main idea behind this approach is to perform two string similarity measurement based on general tokens, correspond to its token sets [18]. If the similarity is denoted, the string pair is flagged as being similar or duplicate. Term-based similarity address drawback on character-based when it works on larger string. In fact, character-based become too computationally expensive and less accurate for imposingly larger strings such as text documents [19]. In this section we will discuss some familiar token-based similarity functions. The main characteristic of token-based similarity is the use of the overlap of two token sets as likeness quantification. The overlap is computed based on exactly matched token pairs without considering other similar tokens. Token-based similarity approach is useful for recognizing the term

rearrangement by breaking the strings into substrings. Jaccard similarity [20], Dice's coefficient [21], Cosine similarity [22], Manhattan distance [23], and Euclidean distance [24] are some examples of these methods.

3.1.2. Corpus-based Similarity

Corpus-based similarity uses a semantic approach. This similarity approach determines the similarity between two concepts based on the information extracted from a large corpora. A corpus (plural corpora) is a large collection of electronic written or spoken text. Corpus contains a predefined set of sentences and their translation to other language. The aim is to match input text with the text in the corpus and achieve translation [25]. Many corpus-based similarity or relatedness measures are based on concept-based resources, such as Wikipedia.

Some of corpus based similarity measures are Hyperspace Analogue to Language (HAL) [26], Latent Semantic Analysis (LSA) [27], Explicit Semantic Analysis (ESA) [28], Pointwise Mutual Information (PMI), Normalized Google Distance (NGD) [29], and Extracting DIstributionally Similar words using CO-occurrence (DISCO) [30].

3.1.3. Knowledge-based Similarity

A semantic similarity measures that uses information from semantic networks to identify the degree of words similarity is called a knowledge-based similarity measures [31]. Knowledge-based similarity consist of semantic similarity and semantic relatedness. Those concepts have been warmly discussed among worldwide researchers. Similarity specifies two interchangeable concepts while relatedness associates concepts semantically [32]. The semantic approach uses an explicit representation of knowledge, such as the interconnection of facts, the meanings of words, and rules to describe conclusions on specific domains. The schema of knowledge representation generally includes the rules of conclusions, logical propositions, and network semantics such as taxonomy and ontology.

Some available ontologies are WordNet, SENSUS1, Cyc2, UMLS3, SNOMED4, MeSH, GO5 and STDS6 [33]. WordNet is the most popular ontology resource and is widely used in knowledge-based similarity measurement. WordNet is a large English lexical database of a research project developed by Princeton University. WordNet organize nouns, verbs, adverbs and adjectives in one concept of semantic relations, called synonym sets (synsets), which represent one concept. Both conceptual-semantic and lexical relations interlinks the sysnets. The words in WordNet are structured hierarchically using hyponymy and hypernym and the words can easily be seen as concepts. In this way, WordNet can be interpreted as a taxonomy. The knowledge-based similarity approach that uses WordNet ontology can be categorized into four measures, path-based, information content-based (IC-based), feature-based, and other types [34].

1) Path-based Measure

The principal concept (also known as edge-counting measures) is the path length and its position in the taxonomy, is represented by a function of similarity between two concepts [35]. This measure uses the shortest length of path between concept, such as the pioneering work of Rada et al. [36], and some of the measures referring to this approach has describe in Lastra-Díaz and García-Serrano [37].

2) IC-based Measure

IC-based approach incorporate a specific concepts in a similarity calculation. The core idea of IC-based similarity measures is applied in an information context (IC) model. The calculation depends on every concept and descendant of frequencies in textual corpus [34]. The fundamental hypothesis should related to the more abstract concept with a lower information rather than a specific content. The IC-based approach is seen as very potential and becomes one of the mainstreams of research in the area so it is still widely discussed in recent years. A novel research was conducted by Lastra-Díaz and García-Serrano [37] who introduced a new ontology-based and new IC-based similarity measures.

3) Feature-based Measure

The main idea of the family of feature-based similarity is using of set-theory operation between concepts feature sets. Feature-based measure describes a set of assumed terms as properties or features. The number of general characteristics are higher than less uncommon characteristics of two terms means that those item are similar [38].

One classical feature-based measure is Tversky's model [39], which argues that similarity is antisymmetric. In between features of subclass and related superclass overcomes the contribution of its inverse direction in terms of similarity evaluation. In recent year, SáNchez and Batet [40] proposed an idea of using the overlapping ancestor sets to estimate the overlapping of unknown feature of the concepts.

3.1.4. Hybrid Similarities

In addition to the three categories previously described, there are still several similarity measures that cannot be categorized into any prior family. The idea of this approach is to combine the previously described approaches, including string-based, corpus-based, and knowledge-based similarity to reach a better metric by adopt their advantages.

Common examples of hybrid metrics are Level2 method proposed by Monge and Elkan [41], SoftTFIDF [42], generalized edit similarity (GES), and Wang et al. [5]. Monge and Elkan [41] propose recursive matching scheme to compare two long string. Implementation of this scheme in which substring are tokens, which call level two distances function. Cohen proposed hybrid metric "soft" TF-IDF similarity use the Jaro-Winkler [13] metric as the "secondary" similarity function. Wang et. al [5] also proposed hybrid similarity function based on token concept, but different from the classical token-based, he employed fuzzy matching between tokens. To quantify the similarity between tokens, this metric uses character-based similarity function.

The most recent hybrid techniques extract semantic knowledge from the structural representation of WordNet and the statistic information on the Internet. Lin [43] proposed a novel linked data (LD) based on hybrid semantic similarity measure, called TF-IDF (LD). The main idea of this algorithm is combine a novel linked data-based TF-IDF scheme with the classical text-based cosine similarity measure. This algorithm integrated in a semi-automatic system (Sherlock) for quiz generation using linked data and textual descriptions of RDF resources. Al-Hasan [44] proposed a new Inferential Ontology-based Semantic Similarity (IOBSS) semantically measure similarity that concern to explicit hierarchical relationship and shared attributes between specific domain items. Atoum and Otoom [45] introduced a novel hybrid on benchmark datasets called text similarity measure (TSM). TSM involves information in WordNet semantic relation such as exactly match words, comparison of sentences pair length, and similarity between word and its reference.

3.2. Experimental Results

The results of the evaluation are shown in Table 1. Test results for the first pair of texts representing two lexically similar terms show that the algorithms in the string-based similarity approach provide a high average score. The Jaro-Winkler and SoftTFIDF algorithms state the highest similarity level with a score of 0.8333. In the semantic-based similarity approach, the highest value is 0.5000 obtained through the Wu Palmer algorithm. Therefore, the text ("book", "cook") pairs have no meaning or semantic relevance.

In the second text pair ("car", "wheel"), almost all of the string-based approaches give a low resemblance value. In a lexical context, these two texts are obviously devoid of character slices. However, the semantic terms "car" and "wheel" are closely related, and the semantic similarity approach expresses the average of high similarity, with the highest value 0.9091 by the Wu Palmer algorithm.

Maximum scores on the third text pairs that represent terms with the highest lexical and semantic proximity are also Wu Palmer's algorithms. In this pair of texts, it appears that the string-based approach also expresses similarities, but more for lexical reasons. In this simple investigation, we can highlight

that the measurement of text similarity using a semantic approach is able to reveal the relatedness between texts

Table 1. Lexical and Semantic Similarity Result

No	Approach/Algorithm	Pair 1	Pair 2	Pair 3
1	Jaro-Winkler	0.8333	0	0.7714
2	N-gram	0.375	1.0	0.5
3	Cosine similarity	0.4999	0	0
4	Jaccard	0.5	0	0.2
5	LSA	0.1485	0.5080	0.1164
6	Wu Palmer	0.5000	0.9091	0.8696
7	Lin	0.1647	0.7355	0.0000
8	Path	0.1429	0.3333	0.2500
9	Monge Elkan	0.7500	0.4000	0.3714
10	SoftTFIDF	0.8333	0	0

4. Conclusion

This article has summarized surveys of measurements of text similarity categorized into four major groups: String-based, Corpus-based, Knowledge-based, and Hybrid similarities. Most common and familiar algorithms in each category have also been reviewed and can be grouped into lexical and semantic similarity. The results of the investigation show that for the purpose of measuring text which emphasizes lexical similarities by ignoring the substance of meaning, the lexical similarity approach is appropriate. These measurements can be used to identify duplication or plagiarism without concern about the document context. String similarity approaches are principally language-independence so they work well for different country languages.

Semantic approach seems to offer intelligence in the measurement of similarities. This measurement is very appropriate to find text or documents that are really similar and conform to the substance of the context. However, the semantic similarities are usually language and domain dependent, so they are not applicable to all languages. In other word, if language ontology is not yet available, it needs to be built first. Referring to the text similarity approaches, it is seen that semantic similarity is very rational to find document similarities. In the future work, our intention is to apply the semantic similarity is applied to the text documents foreshowing the natural relationships among the terms.

References

[1] A. Yunianta, O. M. Barukab, N. Yusof, N. Dengen, H. Haviluddin, and M. S. Othman, "Semantic data mapping technology to solve semantic data problem on heterogeneity aspect," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 3, pp. 161–172, Dec. 2017, doi: https://doi.org/10.26555/ijain.v3i3.131.

[2] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, 2013, doi: https://doi.org/10.5120/11638-7118.

[3] E. Y. Hidayat, F. Firdausillah, K. Hastuti, I. N. Dewi, and A. Azhari, "Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering," *Int. J. Adv. Intell. Informatics*, vol. 1, no. 3, p. 132, Dec. 2015, doi: https://doi.org/10.26555/ijain.v1i3.43.

[4] R. W. Barron and L. Henderson, "The effects of lexical and semantic information on same-different visual comparison of words," *Mem. Cognit.*, vol. 5, no. 5, pp. 566–579, Sep. 1977, doi: https://doi.org/10.3758/BF03197402.

[5] J. Wang, G. Li, and J. Fe, "Fast-join: An efficient method for fuzzy token matching based string similarity join," in *2011 IEEE 27th International Conference on Data Engineering*, 2011, pp. 458–469, doi: https://doi.org/10.1109/ICDE.2011.5767865.

[6] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, Apr. 1950, doi: https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.

[7] V. I. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones," *Probl. Inf. Transm.*, vol. 1, no. 1, pp. 8–17, 1965.

[8] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171–176, Mar. 1964, doi: https://doi.org/10.1145/363958.363994.

[9] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, Mar. 1970, doi: https://doi.org/10.1016/0022-2836(70)90057-4.

[10] R. A. Wagner and M. J. Fischer, "The String-to-String Correction Problem," *J. ACM*, vol. 21, no. 1, pp. 168–173, Jan. 1974, doi: https://doi.org/10.1145/321796.321811.

[11] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, Mar. 1981, doi: https://doi.org/10.1016/0022-2836(81)90087-5.

[12] M. A. Jaro, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *J. Am. Stat. Assoc.*, vol. 84, no. 406, pp. 414–420, Jun. 1989, doi: https://doi.org/10.1080/01621459.1989.10478785.

[13] W. E. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of Record Linkage.," p. 8, 1990, available at: http://files.eric.ed.gov/fulltext/ED325505.pdf.

[14] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the Web," *Comput. Networks ISDN Syst.*, vol. 29, no. 8–13, pp. 1157–1166, Sep. 1997, doi: https://doi.org/10.1016/S0169-7552(97)00031-7.

[15] G. Kondrak, "N-gram similarity and distance," in *International symposium on string processing and information retrieval*, 2005, pp. 115–126, doi: https://doi.org/10.1007/11575832_13.

[16] A. M. Mahdi and S. Tiun, "Utilizing wordnet for instance-based schema matching," in *Proceedings of the International Conference on Advances in Computer Science and Electronics Engineering (CSEE 2014)*, pp. 59–63, available at : http://www.academia.edu/download/34671264/ahmed_CSEE_2014.pdf.

[17] L. Gravano *et al.*, "Approximate string joins in a database (almost) for free," in *VLDB*, 2001, vol. 1, pp. 491–500, available at : http://www.vldb.org/conf/2001/P491.pdf.

[18] M. Yu, G. Li, D. Deng, and J. Feng, "String similarity search and join: a survey," *Front. Comput. Sci.*, vol. 10, no. 3, pp. 399–417, Jun. 2016, doi: https://doi.org/10.1007/s11704-015-5900-5.

[19] M. Y. Bilenko, "Learnable similarity functions and their application to record linkage and clustering," 2006.

[20] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[21] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945, doi: https://doi.org/10.2307/1932409.

[22] A. Bhattacharya, "On a measure of divergence of two multinomial populations," *Sankhya. v7*, pp. 401–406.

[23] E. F. Krause, *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation, 1975.

[24] J. H. Friedman, "On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality," *Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 55–77, 1997, doi: https://doi.org/10.1023/A:1009778005914.

[25] A. Kulkarni, C. More, M. Kulkarni, and V. Bhandekar, "Text Analytic Tools for Semantic Similarity," *Imp. J. Interdiscip. Res.*, vol. 2, no. 5, 2016, available at: http://imperialjournals.com/index.php/IJIR/article/view/688.

[26] K. Lund, "Semantic and associative priming in high-dimensional semantic space," in *Proc. of the 17th Annual conferences of the Cognitive Science Society, 1995*, 1995, pp. 660–665.

[27] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997, doi: https://doi.org/10.1037/0033-295X.104.2.211.

[28] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *IJcAI*, 2007, vol. 7, pp. 1606–1611, available at: http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf.

[29] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google Similarity Distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007, doi: https://doi.org/10.1109/TKDE.2007.48.

[30] P. Kolb, "Disco: A multilingual database of distributionally similar words," *Proc. KONVENS-2008, Berlin*, vol. 156, 2008, available at: http://www.ling.uni-potsdam.de/-kolb/KONVENS2008-Kolb.pdf.

[31] R. Mihalcea, C. Corley, C. Strapparava, and others, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, 2006, vol. 6, pp. 775–780, available at: http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf.

[32] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Comput. Linguist.*, vol. 32, no. 1, pp. 13–47, Mar. 2006, doi: https://doi.org/10.1162/coli.2006.32.1.13.

[33] T. Slimani, "Description and Evaluation of Semantic Similarity Measures Approaches," *Int. J. Comput. Appl.*, vol. 80, no. 10, pp. 25–33, Oct. 2013, doi: https://doi.org/10.5120/13897-1851.

[34] J. J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and F. Chirigati, "HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset," *Inf. Syst.*, vol. 66, pp. 97–118, Jun. 2017, doi: https://doi.org/10.1016/j.is.2017.02.002.

[35] L. Meng, R. Huang, and J. Gu, "A review of semantic similarity measures in wordnet," *Int. J. Hybrid Inf. Technol.*, vol. 6, no. 1, pp. 1–12, 2013, available at: https://pdfs.semanticscholar.org/da95/ceaf335971205f83c8d55f2292463fada4ef.pdf.

[36] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst. Man. Cybern.*, vol. 19, no. 1, pp. 17–30, 1989, doi: https://doi.org/10.1109/21.24528.

[37] J. J. Lastra-D'\iaz and A. Garc'\ia-Serrano, "A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet," 2016, available at: http://e-spacio.uned.es/fez/eserv/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement/Refinement_Espace_LastraGarcia.pdf.

[38] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis, and E. E. Milios, "Semantic similarity methods in wordNet and their application to information retrieval on the web," in *Proceedings of the seventh ACM international workshop on Web information and data management - WIDM '05*, 2005, p. 10, doi: https://doi.org/10.1145/1097047.1097051.

[39] A. Tversky, "Features of similarity.," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977, doi: https://doi.org/10.1037/0033-295X.84.4.327.

[40] T. B. Huedo-Medina, J. Sánchez-Meca, F. Mar'in-Mart'inez, and J. Botella, "Assessing heterogeneity in meta-analysis: Q statistic or I^2 index?," *Psychol. Methods*, vol. 11, no. 2, p. 193, 2006.

[41] A. E. Monge, C. Elkan, and others, "The Field Matching Problem: Algorithms and Applications.," in *KDD*, 1996, pp. 267–270, available at : http://www.aaai.org/Papers/KDD/1996/KDD96-044.pdf.

[42] W. Cohen, P. Ravikumar, and S. Fienberg, "A comparison of string metrics for matching names and records," in *Kdd workshop on data cleaning and object consolidation*, 2003, vol. 3, pp. 73–78, available at : https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf.

[43] C. Lin, D. Liu, W. Pang, and Z. Wang, "Sherlock: A Semi-automatic Framework for Quiz Generation Using a Hybrid Semantic Similarity Measure," *Cognit. Comput.*, vol. 7, no. 6, pp. 667–679, Dec. 2015, doi: https://doi.org/10.1007/s12559-015-9347-7.

[44] M. Al-Hassan, H. Lu, and J. Lu, "A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system," *Decis. Support Syst.*, vol. 72, pp. 97–109, Apr. 2015, doi: https://doi.org/10.1016/j.dss.2015.02.001.

[45] I. Atoum and A. Otoom, "Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 9, pp. 124–130, 2016, doi: 10.14569/IJACSA.2016.070917, available at :http://thesai.org/Publications/ViewPaper?Volume=7&Issue=9&Code=ijacsa&SerialNo=17.